# 新聞画像のレイアウト解析に関する研究

佐々木浩士*・大倉　充**・塩野　充**

*岡山理科大学大学院工学研究科修士課程情報工学専攻

**岡山理科大学工学部情報工学科

（1998年10月5日　受理）

## 1．まえがき

情報化社会の現代において，情報を入手する方法には様々なものがある。その例として，新聞や雑誌，TV，ラジオ等が挙げられる。最近ではコンピュータネットワークからの情報の入手が盛んになりつつあり，このため新聞を代表とする紙で流通していた情報が，電子媒体を通して利用されることが多くなってきている。その新聞などの文書をコンピュータ上で作成する場合には，ワードプロセッサなどのハードウェア的な支援環境の発展は著しく，それらを利用し多くの人が容易に印刷文書並の文書を作成できるようになっている。特に，様々なエディタなどのソフトウェア的な処理環境は着実に発展している。ところが，作成された文書の利用（読む）のための支援環境に関しては，種々の OCR が開発されているが，まだまだ研究段階にあると考えられる[1,2]。そこで本研究では，新聞画像を対象としてそこから記事中のキーとなる情報を自動的に抽出し分類するための，レイアウト解析について基礎的な検討を行う[3]。

## 2．使用データ

本研究では，工業技術院電子技術総合研究所作成の文書画像データベース JEIDA93に含まれる，以下に示す4つのデータを使用した。

朝日新聞　第一面　600 dpi　12925 × 9512 pixels

朝日新聞　経済面　600 dpi　13071 × 9504 pixels

毎日新聞　第一面　600 dpi　13052 × 9520 pixels

毎日新聞　経済面　600 dpi　13057 × 9544 pixels

これらのデータは非常に大きいため，必要に応じて縮小して処理を行っている。図1に使用した朝日新聞第一面の新聞画像を示す。

## 3．解析における問題点

一般的に，新聞はページごとに異なったレイアウト構造を持っている。記事の構成にしても，見出しや文書，写真，表，グラフ等様々な領域を含んでいる。そのためこれらの領


図1 朝日新聞 第一面


図2 新聞画像の特徴

域の正確な抽出及び分類が必要である。また本研究で用いた新聞画像には広告の領域が含まれているために,この領域と上述した記事を構成する領域との区別が必要となってくる。広告の多くは新聞の下部に位置するが,記事内に含まれるものも存在するため,その区別は容易ではない。図2に新聞画像の特徴を示す。

## 4．レイアウト解析の概要

本研究におけるレイアウト解析では,新聞画像に対して(1)外接矩形の抽出,(2)各種領域の判定,(3)レイアウト状況の表示の順序で処理を行う。以下,各処理について説明を行う。

### 4．1 外接矩形の抽出

入力された新聞画像に対して,雑音除去処理を行いその画像から8連結の黒画素に対してラベリング処理[4]を行う。そしてラベル付けされた連結成分に外接する矩形を求める。方形領域の作成は図3に示すように,連結している黒画素の横方向の最小点 sx,最大点 ex,縦方向の最小点 sy,最大点 eyを求めることによって行う。作成された方形領域から,矩形データとして図3と表1に示す左上のx座標,右下のx座標,左上のy座標,右下のy座標,矩形の高さ,幅,面積,矩形中の画素数という8種類のデータを取得する。後述の処理は,極力この矩形データにアクセスすることにより処理の高速化を計る。図4に朝日新聞第一面の外接矩形の抽出結果を示す。抽出される矩形数は約6000であり,図4の画像の場合は矩形数6227である。

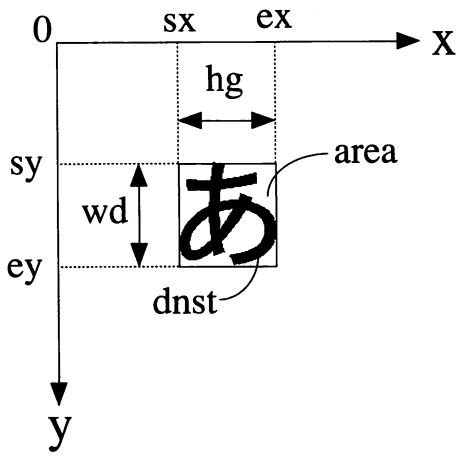


図3　外接矩形特徴

表1　矩形データ

| 特徴名 | 記号 |
|---|---|
| 左上のx座標 | sx |
| 右下のx座標 | ex |
| 左上のy座標 | sy |
| 右下のy座標 | ey |
| 矩形の幅 | wd |
| 矩形の高さ | hg |
| 矩形の面積 | area |
| 矩形中の面素数 | dnst |

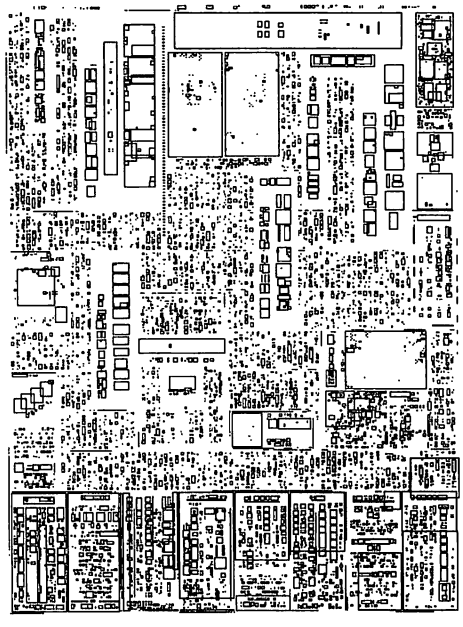

図4　矩形抽出結果（ラベル数6227）

4．2　各種領域の判定

前処理によって入手した外接矩形のデータを基に，直線，広告，写真，見出しなどの各領域ごとの特徴を利用して抽出を行う。

4．2．1　直線領域

新聞画像に限らずレイアウト解析を行う場合，文書を区分する直線は非常に有効なものである。しかし本研究では，画像の縮小方法として直線を重視せずに他の見出しや写真などが鮮明に表示される方法を選択している。そのため，縮小した画像中には有効な

直線部分がほとんど失われている。このため直線の領域は不必要なものと判断し，矩形データより高さまたは幅の値が0になる領域を直線として除去する。また直線だけでなく罫線も途中で途切れているものが存在するため，矩形の面積と矩形中の画素数を比較することにより除去を行う。

4.2.2 下部の広告領域

新聞には，紙面の下部に広告が含まれることが多い。広告が含まれている場合には，通常，記事部分とは直線によって区分されている。直線で区分されていない場合においても，明らかに記事の部分とはある程度の間隔が設けられている。このことを利用して下部の広告領域の抽出を行う。まず図5に示す直線領域を除去した矩形データに対して，横一列が全て白画素の部分の有無を調査する。横一列が全て白画素の部分がある程度連続する場合には，その部分を記事と広告領域を区分する空白と見なし，その空白部分から下部を広告領域と決定する。さらにその広告領域は何種類かに分かれている場合があるので，その領域中の縮小データから縦方向の周辺分布を求める。その周辺分布の状況によって，広告領域内を分割する。図6には，広告領域中の周辺分布によって，その座標に存在する画素数が0になる部分を示している。

4.2.3 写真領域

新聞に含まれる写真は，主にそのページの中で重要な項目に付属して比較的大きく扱われている。矩形データからも，矩形の高さや幅が一定の大きさを持ち矩形内の黒画素数もかなり多いことがわかる。このことを利用して写真領域の抽出を行う。まず矩形の

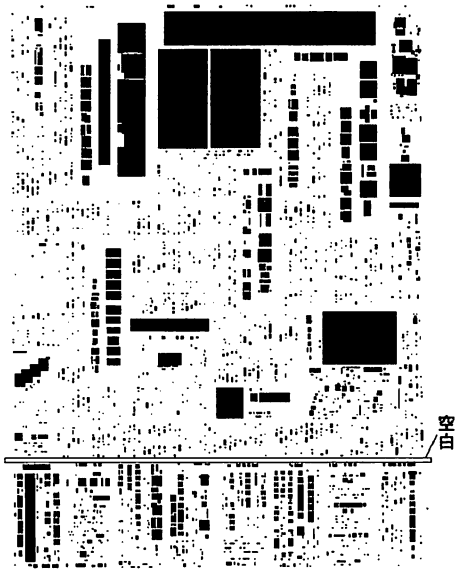


図5　直線領域を除去した画像

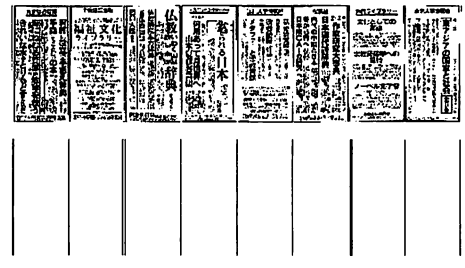

図6　広告領域中の分割位置

データから矩形の高さや幅が一定値以上の大きさを持ち，矩形内部の黒画素の数が面積の半分以上のものを写真領域の候補として選択する。選択された矩形データに対して，その矩形中の縮小データから，矩形の高さまたは幅で長い側を基準とした周辺分布を求める。このときの周辺分布は黒画素に対するものではなく，後述の白抜き見出し領域と区別するために，白画素に対しての周辺分布を用いる。図7に示すように，周辺分布の状態が連続的，つまり白画素のヒストグラムが急激な増減をしておらず，矩形の高さと幅の比率が一定値以下であれば写真領域として抽出する。

4．2．4　白抜き見出し領域

白抜き見出しとは，黒を下地にして白抜きの文字で書かれている見出しのことである。そのページの中で最も重要な項目に付けられることが多く，写真と同様に大きく扱われる。また見出しは文字によって構成されているため，幅か高さのどちらかの方向に細長くなり偏る傾向がある。矩形データからも，矩形の高さや幅が一定の大きさを持ち，高さと幅の比率が大きく矩形内の黒画素数もやや多いことがわかる。このことを利用して白抜き見出し領域の抽出を行う。まず矩形のデータから矩形の高さや幅が一定値以上の大きさを持ち，矩形内部の黒画素の数が面積の半分以上のものという写真領域と同様の条件により，白抜き見出し領域の候補を選択する。選択された矩形データに対して，その矩形中の縮小データから，矩形の高さまたは幅で長い側を基準とした周辺分布を求める。このときの周辺分布も，写真領域と同様に白画素に対しての周辺分布を用いる。図8に示すように，周辺分布の状態が断続的，つまり白画素のヒストグラムが急激な増減をしており，矩形の高さと幅の比率が一定値以上ならば白抜き見出し領域として抽出する。

4．2．5　見出し領域

白抜き見出し以外の見出しは，文字の大きさによって様々な種類がある。しかも外接矩形の抽出の際には一文字ごとに抽出されるため，それらの統合が必要になってくる。

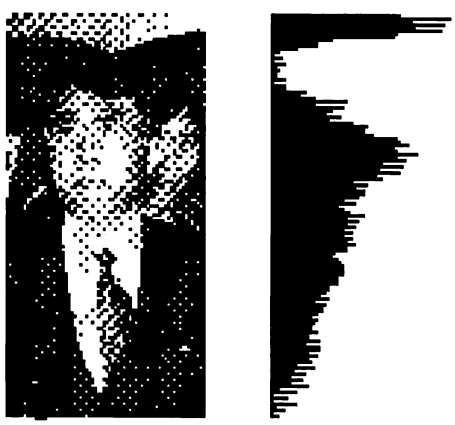

図7　写真領域の周辺分布


図8　白抜き見出し領域の周辺分布

このため上述の処理によって抽出された領域と区別するために，領域が確定した矩形を除去したデータを作成する。このデータに対して膨張収縮処理を行い矩形データを再構成する。膨張収縮処理の例を図9に示す。再構成された矩形データから矩形の高さや幅がある程度の大きさをもつものを見出し領域の候補として選択する。選択された矩形データに対して，その矩形中の縮小データから，矩形の高さまたは幅で長い側を基準とした周辺分布を求める。この周辺分布は，普通の見出しに対するものであるため，黒画素に対しての周辺分布を用いる。図10に示すように，周辺分布の状態が断続的，つまり黒画素のヒストグラムが急激な増減をしており，矩形の縦横比または横縦比が一定値以上ならば見出し領域として抽出する。

### 4．3　各種領域の表示

各々の判定された領域に対して，区別するために色分けしてレイアウト状況を表示する。

## 5．解析結果

図11に朝日新聞第一面の解析結果を示す。また表2にデータ別の抽出数を，表3に処理時間を示す。抽出数は，抽出された領域/画像中に存在する領域数で表している。その他の領域は，見出し領域の抽出の際に候補として選択されたが，見出し領域としては決定されなかったものである。解析は前述の4つのデータに対して行った。写真領域は，一定の大きさを持つ領域は抽出可能であるが，人物の顔写真などの小さな領域を持つものが抽出できていない。これらの領域は，図やグラフ，広告領域である可能性もあるため，抽出が困


図9　膨張収縮処理の例


図10　見出し領域の周辺分布

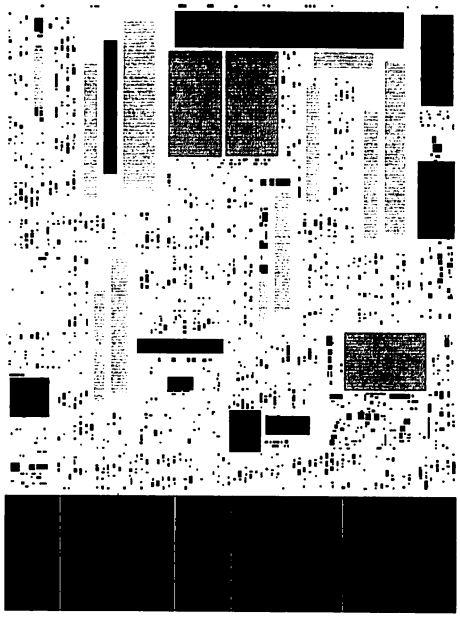

図11　レイアウト解析結果

表2　データ別の抽出数

| | 写真領域 | 白抜き見出し領域 | 見出し領域 | 広告領域 | その他の領域 |
|---|---|---|---|---|---|
| 朝日第一面 | 3/3 | 3/3 | 9/12 | 8/12 | 5 |
| 朝日経済面 | 0/1 | 0/1 | 12/16 | 2/3 | 2 |
| 毎日第一面 | 2/5 | 1/1 | 8/9 | 6/10 | 11 |
| 毎日経済面 | 1/2 | 1/1 | 10/15 | 2/2 | 4 |
| 合　計 | 6/11 | 5/6 | 39/52 | 18/27 | 22 |

表3　データごとの処理時間

| データ名 | 処理時間（秒） |
|---|---|
| 朝日新聞第一面 | 282.43 |
| 朝日新聞経済面 | 311.48 |
| 毎日新聞第一面 | 295.52 |
| 毎日新聞経済面 | 329.51 |

難となっている。白抜き見出し領域は，白抜き見出しと普通の見出しが重なっているものを普通の見出しとして抽出したものが一つ存在した。広告領域は，下部に存在するものは全て抽出できているが，記事中に含まれる広告領域は抽出できていない。広告自体に様々な種類があり，その特徴を限定することは難しいためである。見出し領域は，文字サイズが大きなものは膨張収縮処理によって矩形の統合が容易だが，文字サイズの小さいものやひらがなや数字など，矩形のサイズが小さくなるような文字が多く含まれる見出しなども

存在し，それらに膨張収縮処理を行ってもあまり効果がないため抽出が不完全となっている。

## 6．むすび

本研究では，新聞画像のレイアウト解析に関する検討を行った。写真領域と見出し領域に関しては，領域がある程度の大きさを有する場合には抽出可能である。しかし，記事に含まれる広告や図・グラフなどに関しては，現状の方法では抽出が不完全である。また文書領域の統合もできていないため，それらを検討することも課題の一つである。

### 参考文献

1）駱　琴，渡辺豊英，杉江　昇：ルールベース適用による日本語新聞紙紙面の構造認識，信学論(D-II)，vol. J75-D-II，no. 9，pp. 1514-1225 (1992)．
2）平山唯樹：複雑なカラム構造をもつ文書イメージの領域分割法，信学論(D-II)，vol. J79-D-II, no. 11, pp. 1790-1799 (1996)．
3）佐々木浩士，大倉　充，塩野　充：新聞画像のレイアウト解析，電気・情報通信学会中国支部第48回連合大会，072222，p. 207 (1997年10月)．
4）長谷川純一，興水大和，中山　晶，横井茂樹：画像処理の基本技法，技術評論社，(1986)．

# Layout Analysis of Japanese Newspapers

Kouji SASAKI*, Mitsuru OHKURA** and Mitsuru SHIONO**

*Graduate School of Engineering*

***Department of Information and Computer Engineering*

*Okayama University of Science,*

*Ridai-cho 1-1, Okayama 700-0005, Japan*

(Received October 5, 1998)

Concerning today of an information-intensive society, there are various ways to get information-for example, the newspaper, the magazine, the TV, the radio and the computer network. Especially the network is about to be popular in these days. From this reason, many pieces of information can be obtained from the network and the electronic medium. When we make a document on a computer, either the hardware such as a word processor or a computer software may be used. Such a development of support environment for making a document is marked, however, the utilization of publications on a computer has still problem in spite of inventing the OCR. In this paper, a basic examination is performed about the layout analysis of Japanese newspapers to make a database automatically. The items to be extracted from the papers are the headline, the photograph, the drawing and the advertisement.